MULTINET: A BENCHMARK FOR GENERALIST ACTION MODELS

A PREPRINT

David S. Hippocampus*
Department of Computer Science
Cranberry-Lemon University
Pittsburgh, PA 15213
hippo@cs.cranberry-lemon.edu

Elias D. Striatum
Department of Electrical Engineering
Mount-Sheikh University
Santa Narimana, Levand
stariate@ee.mount-sheikh.edu

November 6, 2024

ABSTRACT

Recent advances in machine learning have demonstrated the potential of large-scale models to exhibit broad generalization capabilities across diverse tasks. However, the evaluation of these models remains fragmented, with different benchmarks focusing on specific modalities or capabilities. We present Multinet, a comprehensive benchmark designed to evaluate truly generalist models across vision, language, and action domains. Multinet consolidates diverse, high-quality datasets including OBELICS, COYO-700M, and OpenX-Embodiment, establishing standardized evaluation protocols for assessing both the action capabilities of Vision-Language Models (VLMs) and the multimodal understanding of Vision-Language-Action Models (VLAs). Our benchmark includes carefully curated training data spanning vision-language association (800M+ image-text pairs), language understanding (1.3T tokens), and control tasks (35+ TB of robotics and RL data). We provide evaluation protocols across multiple dimensions, including image captioning, visual question answering, commonsense reasoning, and robotic control. Additionally, we open-source a toolkit that standardizes the challenging process of obtaining and utilizing reinforcement learning and robotics data from various sources. Through systematic evaluation of state-of-the-art models, we demonstrate significant gaps in current approaches: VLMs struggle with control tasks while VLAs show limited capabilities in pure vision-language understanding. These findings highlight the need for more genuinely generalist models. Multinet serves as both a comprehensive evaluation framework and a foundation for developing the next generation of truly generalist AI systems.

1 Introduction

Motivations

Recent advances in machine learning have demonstrated the potential of large-scale models to exhibit broad generalization capabilities across diverse tasks. A particularly promising direction emerged with DeepMind's Gato [**?**], which provided the first glimpse of a truly generalist model capable of performing hundreds of different tasks across multiple domains. However, the datasets and model used to train Gato remain closed-source, limiting the research community's ability to build upon and extend this work.

Concurrent developments in Vision-Language-Action (VLA) models have shown impressive capabilities in grounding natural language instructions in real-world actions [**? ?**]. These models can interpret visual scenes, understand language commands, and generate appropriate control actions. However, current VLA models are primarily focused on narrow domains like robotic manipulation, and often struggle with more complex vision-language understanding tasks. This limitation stems partly from their training data, which typically emphasizes a specific subset of capabilities rather than true generalist behavior.

*Use footnote for providing further information about author (webpage, alternative address)—*not* for acknowledging funding agencies.

Building genuinely generalist models requires training on diverse datasets that span multiple modalities and task types. Such models must excel not only at individual modalities (vision, language, or control) but also at tasks that require seamless integration across modalities - a requirement that better reflects real-world scenarios. Currently, there exists no large-scale, open-source dataset specifically designed for training and evaluating such generalist models. This gap motivated the development of Multinet.

Contributions

In this paper, we present Multinet, a comprehensive benchmark for developing and evaluating generalist action models. Our contributions include:

- The largest open-source generalist dataset, consolidating diverse modalities and tasks suitable for pre-training, fine-tuning, and evaluation
- A novel benchmark for assessing state-of-the-art Vision-Language Models (VLMs) and Vision-Language-Action Models (VLAs)
- A detailed analysis of the constituent datasets' validity and utility for generalist objectives
- An open-source software toolkit that facilitates dataset access and standardizes control data from various sources into a common TensorFlow format

Through Multinet, we aim to accelerate research in generalist models by providing the community with the necessary tools and benchmarks to develop and evaluate truly general-purpose AI systems. Our benchmark enables systematic comparison of different approaches and provides insights into the challenges and opportunities in building models that can seamlessly operate across multiple modalities and tasks.

2 Related Work

Recent advances in machine learning have produced several promising directions toward generalist AI systems. We organize our discussion of related work into three main categories: generalist agents, vision-language-action models, and existing benchmarks.

2.1 Generalist Agents

DeepMind's Gato [**?**] represented a significant milestone as the first truly multi-modal, multi-task, multi-embodiment agent. Operating with a single neural network and fixed weights, Gato demonstrated capabilities spanning Atari gameplay, image captioning, conversational interaction, and real-world robotic manipulation. The agent dynamically selects appropriate output modalities (text, joint torques, button presses, or other tokens) based on context. However, both the model and the majority of its training datasets remain closed-source, limiting its impact on the broader research community.

2.2 Vision-Language-Action Models

Recent work has produced several notable vision-language-action (VLA) models, including Octo [**?**] and OpenVLA [**?**]. These models demonstrate impressive capabilities in grounding language instructions in robotic control actions through either pre-training from scratch or fine-tuning existing models. However, their training focuses exclusively on vision-language-grounded control tasks, neglecting pure vision-language understanding or generation tasks. This specialization limits their ability to serve as truly generalist models capable of operating across the full spectrum of modalities.

2.3 Benchmarks and Evaluation Frameworks

Existing benchmarks in the field can be categorized into control-focused, vision-language focused, and robotics-specific evaluations.

Control Benchmarks: RL Unplugged [**?**] provides a comprehensive suite of offline reinforcement learning benchmarks, spanning domains from Atari games to DM Control Suite tasks. It standardizes environments, datasets, and evaluation protocols to enhance reproducibility in offline RL research. Similarly, D4RL [**?**] offers over 40 standardized environments and datasets for offline RL, covering robotic manipulation, navigation, and autonomous driving tasks. The robotics community has developed numerous specialized benchmarks, particularly focusing on imitation learning and behavior cloning. THE COLOSSEUM provides a systematic evaluation framework for robotic manipulation across 14

environmental perturbations, while FactorWorld and KitchenShift examine generalization across various environmental factors. Several task-specific benchmarks have emerged: RLBench offers 100 simulated manipulation tasks, RAVENS focuses on vision-based manipulation, and FurnitureBench addresses long-horizon complex manipulation in real-world settings. Recent additions include LIBERO for lifelong robot learning, FMB for generalizable manipulation, DUDE for document manipulation, and ProcTHOR for procedurally generated embodied AI tasks.

Vision-Language Benchmarks: The evolution of multimodal evaluation has progressed significantly from single-task benchmarks like VQA, OK-VQA, and MSCOCO to more comprehensive frameworks. MultiBench [**?**] presents a unified evaluation framework spanning 15 datasets, 10 modalities, and 20 prediction tasks across 6 research areas. MMLU [**?**] evaluates language model capabilities across 57 academic subjects, while MMMU extends this to college-level multimodal understanding across technical disciplines. Recent developments include specialized benchmarks like MathVista for mathematical reasoning and GAIA for fundamental reasoning abilities, as well as more holistic evaluations through LAMM, LVLM-eHub, SEED, and MM-Vet. LiveBench [**?**] specifically addresses test set contamination through continuously updated evaluation data.

While these benchmarks excel in their respective domains, they remain specialized to particular modalities or task types. Multinet addresses this limitation by providing a diverse collection of datasets specifically designed to evaluate and advance truly generalist models. Our benchmark enables the development of systems with strong capabilities across vision-language association, language understanding and generation, and reward-based action trajectories in varied environments. Additionally, Multinet serves as a comprehensive evaluation framework for current VLMs and VLAs, highlighting areas for improvement in the development of next-generation generalist models.

3 Coverage

3.1 Datasets

3.1.1 For Training

Vision-Language and Language

OBELICS OBELICS [**?**] is an open web-scale filtered dataset of interleaved image-text documents extracted from Common Crawl [**?**]. It comprises 141 million web pages, 353 million associated images, and 115 billion text tokens. The interleaved nature of the documents provides richer context compared to simple image-text pairs. The dataset occupies 666 GB in arrow format and 377 GB in Parquet format, and currently includes only training data.

COYO-700M COYO-700M [**?**] contains 747 million pairs of alt-text and associated images harvested from HTML documents. The dataset was curated from approximately 10 billion initial pairs collected from CommonCrawl between October 2020 and August 2021. It employs minimal filtering, resulting in a "noisier" dataset that can potentially improve model robustness while providing challenging evaluation scenarios. The dataset occupies 135 GB before image conversion.

MS-COCO Captions MS-COCO Captions [**?**] is a large-scale object detection, segmentation, and captioning dataset containing 330,000 images with 5 captions each, totaling 1.5 million object instances. The dataset's high-quality annotations make it particularly valuable for training and evaluating vision understanding tasks.

Conceptual Captions Conceptual Captions [**?**] consists of 3.3 million web-harvested images with filtered descriptions derived from HTML alt-text attributes. The dataset includes 3,318,333 image-URL/caption pairs for training and 15,840 pairs for validation. Unlike COYO-700M, it undergoes more rigorous filtering to ensure higher data quality.

A-OKVQA A-OKVQA [**?**], the successor to OKVQA [**?**], contains 24,903 question/answer/rationale triplets requiring broad commonsense and world knowledge. The dataset is split into 17.1K/1.1K/6.7K for train, validation, and test. Questions cannot be answered by simply querying a knowledge base, making it particularly valuable for evaluating sophisticated reasoning capabilities.

VQA-V2 VQA-V2 [**?**] provides open-ended questions about images that require understanding of vision, language, and commonsense knowledge. The dataset includes 265,000 images with at least three questions per image, ten ground truth answers per question, and three plausible answers per question.

DataComp-1B DataComp-1B [**?**] consists of 1.4 billion image-text pairs curated from an initial pool of 12.8 billion pairs. Despite being smaller than alternatives like LAION-2B [**?**], it achieves better performance with fewer resources, demonstrating 79.2

Fineweb-edu Fineweb-edu [**?**] contains 1.3T tokens of educational content filtered from the FineWeb dataset using an educational quality classifier trained on LLama3-70B-Instruct [**?**] annotations. The dataset demonstrates superior performance on standard benchmarks compared to larger, unfiltered alternatives.

DM Lab DM Lab [**?**] provides frames from the DeepMind Lab environment annotated with agent-object distances. The dataset contains 360x480 color images across 6 classes (combinations of close, far, very far and positive reward, negative reward). This 1.8 TB dataset is particularly valuable for training models to reason about spatial relationships and depth perception in 3D environments, making it crucial for robotics and augmented reality applications.

ALE Atari The Arcade Learning Environment (ALE) provides 57 Atari 2600 game environments for AI agent development. The dataset version included in Multinet contains 500,000 interactions per game generated by JAT [**?**], where dedicated agents were trained for 2 billion steps using asynchronous PPO [**?**]. The complete dataset occupies 66 GB.

BabyAI BabyAI [**?**] is a research platform comprising 19 levels of increasing difficulty, designed to investigate grounded language learning with humans in the loop. The platform teaches agents a combinatorially rich subset of English and includes a hand-crafted bot that simulates a human teacher. Our version contains 100,000 episodes across 39 available settings collected by JAT [**?**], totaling 148 GB.

MuJoCo The MuJoCo [**?**] benchmark suite contains 11 continuous control tasks of varying complexity. Our dataset includes 10,000 episodes per environment, generated by agents trained using asynchronous PPO [**?**] from Sample Factory [**?**]. These agents achieved scores meeting or exceeding current standards. The complete dataset occupies 33 GB.

DM Control Suite The DeepMind Control Suite [**?**] provides standardized continuous control environments powered by the MuJoCo physics engine. It includes diverse tasks ranging from simple (Pendulum, Cart-pole) to complex (Humanoid, Manipulator), with interpretable rewards. The dataset occupies 52 GB and is particularly valuable for training on simulated motor-control problems.

V-D4RL V-D4RL [**?**] is the first publicly available benchmark for continuous control from visual observations of DMControl Suite tasks featuring diverse behavioral policies. The 62 GB dataset specifically tests robustness to distractions, generalization across dynamics, and offline reinforcement learning at scale.

Meta-World Meta-World's MT50 benchmark [**?**] provides 50 diverse robot manipulation tasks. Our 15 GB dataset contains 10,000 episodes per environment (limited to 100 timesteps) generated by task-specific trained agents. The agents successfully solved most tasks except Assembly and Disassemble.

Procgen Procgen [**?**] is OpenAI's suite of 16 procedurally generated game-like environments designed to benchmark efficiency and generalization in RL. With diversity comparable to ALE, these environments require robust policies that avoid overfitting to narrow state spaces. The dataset occupies 739 GB.

OpenX-Embodiment OpenX-Embodiment [**?**] is currently the largest open-source real robot dataset, containing over 1M trajectories from 22 robot embodiments. For Multinet v0.1, we utilize 53 of the 72 available datasets, stored in the RLDS [**?**] format which accommodates various action spaces and input modalities. The training splits of these 53 datasets total 32 TB. Dataset selection involved careful curation based on robot morphology, gripper specifications, action spaces, and sensor configurations.

LocoMuJoCo LocoMuJoCo [**?**] is an open-source imitation learning benchmark focused on locomotion. It includes diverse environments (quadrupeds, bipeds, and musculoskeletal human models) with comprehensive datasets including real noisy motion capture data and ground truth expert data. Our version uses the "perfect dataset" containing ground truth states and actions from expert policies, occupying 690 MB.

3.1.2 For Evaluation

Beyond the validation and test splits available in our training datasets, we include several datasets specifically for evaluation purposes:

Flickr30k Flickr30k [**?**] contains 31,000 images from Flickr, each paired with five human-annotated descriptive sentences. The dataset includes rich annotations like coreference chains and bounding boxes, making it particularly valuable for evaluating sentence-based image description and grounded language understanding. The dataset occupies 4.1 GB.

TextVQA TextVQA [**?**] evaluates models' ability to read and reason about text within images. It contains 45,336 questions across 28,408 images from OpenImages [**?**], specifically testing models' capability to incorporate text as a modality for question answering.

VizWiz VizWiz [**?**] provides a collection of images taken by blind individuals along with associated questions, presenting unique challenges due to poor image quality and conversational question formats. This dataset tests model robustness in real-world assistive technology scenarios where questions may not always be answerable.

WinoGAVil WinoGAViL [**?**] tests vision-and-language commonsense reasoning abilities by requiring models to identify associations beyond simple visual recognition. It provides both zero-shot and supervised evaluation settings, challenging models with tasks requiring general knowledge and abstraction capabilities.

ImageNet-R ImageNet-R [**?**] contains artistic renditions of 200 ImageNet [**?**] classes, including art, cartoons, graffiti, embroidery, and other stylized representations. The dataset evaluates models' ability to generalize beyond standard photographic images to diverse visual representations.

ObjectNet ObjectNet [**?**] is a real-world test set for object recognition featuring random backgrounds, rotations, and viewpoints. Without an associated training set, it specifically tests generalization capabilities by controlling for common dataset biases and presenting objects in unusual poses and cluttered scenes.

HellaSwag HellaSwag [**?**] comprises 70,000 multiple-choice questions designed to evaluate commonsense natural language inference. The dataset features adversarially generated endings that are difficult for AI but easy for humans, who achieve over 95

WinoGrande WinoGrande [**?**] contains 44,000 examples inspired by the Winograd Schema Challenge [**?**]. Each example presents a sentence with an ambiguous pronoun that requires commonsense reasoning and world knowledge to resolve correctly.

ARC The AI2 Reasoning Challenge [**?**] provides multiple-choice questions from grade 3-9 science exams, split into Easy and Challenge sets. The Challenge Set specifically tests advanced reasoning capabilities that many current algorithms struggle with.

CommonsenseQA CommonsenseQA [**?**] contains 12,247 multiple-choice questions, each with one correct answer and four distractors. The dataset evaluates models' ability to utilize general world knowledge beyond specific contexts.

MMLU The Massive Multitask Language Understanding benchmark [**?**] evaluates language models across 57 subjects through approximately 16,000 multiple-choice questions. It tests world knowledge and problem-solving abilities in zero-shot and few-shot settings across diverse academic disciplines.

3.2 Analysis

Multinet consists of datasets across several categories, encompassing a diverse set of modalities and tasks, thus testing a potential generalist model in many different ways. **All control datasets are simulated environments except for the OpenX-Embodiment collection which contains data from robots used in the real-world.**

Dataset	Task/Data type	Modality
OBELICS	Interleaved Image-Text	Vision-Language
COYO-700M	Image-Text pairs	Vision-Language
MS COCO	Object detection, segmentation, key-point detection, captioning	Vision-Language
Conceptual Captions	Image Captioning	Vision-Language
A-OKVQA	Visual Question Answering	Vision-Language
VQA-v2	Visual Question Answering	Vision-Language
Datacomp-1B	Image-Text pairs	Vision-Language
Fineweb-edu	High quality text corpus	Language

Table 1: Training, fine-tuning, and evaluation datasets for vision-language and language tasks

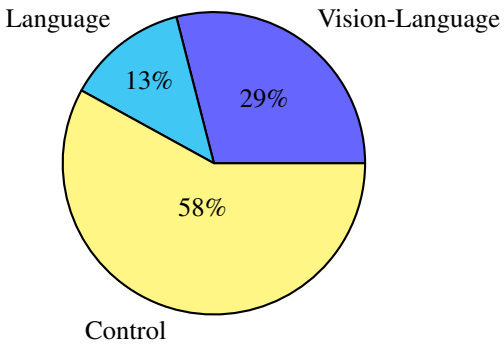

Figure 1: Distribution of datasets across modalities in Multinet. Control represents the largest portion (58%) due to the extensive OpenX-Embodiment collection, followed by Vision-Language (29%) and Language (13%) datasets.

4 The Multinet Benchmark

4.1 Motivation

While existing benchmarks excel at evaluating specific capabilities and modalities, there remains a notable gap in holistic evaluation frameworks that can assess both the action capabilities of Vision-Language Models (VLMs) and the multimodal understanding of Vision-Language-Action Models (VLAs). Multinet addresses this gap by providing a comprehensive benchmark that spans vision-language, language, and control tasks. Our work consolidates diverse, high-quality datasets and establishes standardized evaluation metrics to enable systematic comparison of state-of-the-art models.

4.2 Evaluation Metrics

We employ several complementary metrics to evaluate model performance across different modalities and tasks:

CIDEr The Consensus-based Image Description Evaluation (CIDEr) metric [**?**] evaluates image captioning quality by comparing generated captions to reference captions using TF-IDF-weighted n-grams. For a candidate caption c_i and reference captions S_i, the CIDEr score for n-grams of length n is calculated as:

$$\text{CIDEr}_n(c_i, S_i) = \left(\frac{1}{m} \sum_j \text{sim}(c_i, s_{ij}) \right) \tag{1}$$

where similarity is computed as:

$$\text{sim}(c_i, s_{ij}) = \frac{g^n(c_i) \cdot g^n(s_{ij})}{||g^n(c_i)|| \cdot ||g^n(s_{ij})||} \tag{2}$$

The final CIDEr score averages across different n-gram lengths:

$$\text{CIDEr}(c_i, S_i) = \frac{1}{N} \sum_{n}^{N} \text{CIDEr}_n(c_i, S_i) \tag{3}$$

VQA Accuracy This metric evaluates visual question answering performance through exact string matching between predicted and reference answers:

$$\text{VQA accuracy} = \frac{\sum_{i=1}^{N} \min\left(1, \frac{\text{number of annotators who agree with model's answer}}{3}\right)}{N}$$

Recall@K For image-text retrieval tasks, Recall@K measures the proportion of relevant items retrieved within the top K results:

$$\text{Recall@K} = \frac{\text{Number of relevant items retrieved within top K}}{\text{Total number of relevant items}}$$

Accuracy For commonsense reasoning and text understanding tasks, we use standard classification accuracy:

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}} \times 100 \tag{4}$$

Mean Squared Error For evaluating action prediction on offline robotics trajectories, we employ Mean Squared Error:

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{5}$$

MSE is particularly suitable for this task due to its non-negativity, sensitivity to large errors, and incorporation of both bias and variance components. When evaluating models on the OpenX dataset, we use MSE to measure the accuracy of predicted actions given observation states, image observations, and language instructions at each timestep. This metric is especially valuable for offline evaluation where direct robot deployment is not possible.

4.3 Benchmark Evaluation Protocol

The first release of Multinet establishes evaluation protocols for state-of-the-art models including GPT-4, JAT, and OpenVLA on the OpenX-Embodiment datasets [**?**]. Our evaluation strategy adapts to the varying availability of dataset splits:

Pre-existing Test Splits For datasets with established test splits, we conduct evaluations directly on these splits, maintaining consistency with previous benchmarking efforts.

Validation Split Usage When test splits are unavailable but validation splits exist, we use the validation splits for evaluation to avoid potential training data contamination.

Custom Split Creation For datasets containing only training data, we create evaluation splits by:

- Reserving approximately 20% of the dataset shards for evaluation
- Ensuring each shard contains 512 timesteps of data
- Extending the evaluation split boundary when necessary to include complete episodes (i.e., if an episode extends beyond the 20% boundary, we include all timesteps until episode completion)

Metric	Evaluation Categories
CIDEr	• Image Captioning • Image-based Text Retrieval
VQA Accuracy	Visual Question Answering
Recall@K	• Image Understanding • Text-based Image Retrieval
Accuracy	• Visual Question Answering • Commonsense Reasoning • Text Understanding
Mean Squared Error	• Reinforcement Learning • Robotics

Table 2: Metrics used in the Multinet benchmark and their corresponding evaluation categories. Each metric is designed to assess specific aspects of model performance across different modalities and tasks.

This protocol ensures comprehensive evaluation while maintaining episode integrity and preventing the artificial truncation of behavioral sequences.

5 Importance of Multinet

The development and release of Multinet represents a significant step toward advancing generalist AI systems. Its importance spans several key dimensions:

Advancing Generalist Foundation Models Multinet establishes a comprehensive benchmark for evaluating truly generalist models that can operate across multiple modalities, tasks, and environments. Our initial findings [] demonstrate a significant capability gap in current state-of-the-art models: while VLMs and VLAs excel in their primary domains, they struggle to maintain consistent performance across vision-language, language, reinforcement learning, and robotics tasks.

Enabling Next-Generation VLA Models Current Vision-Language-Action models typically excel at vision-language-grounded actions but may underperform in pure vision-language or language tasks. Multinet provides pre-training scale data across all these modalities, enabling the development of models that achieve state-of-the-art performance across all constituent tasks, not just their primary domain. As future versions expand the dataset, these capabilities will only grow stronger.

Comprehensive Evaluation of Robotics Foundation Models Traditional evaluations of robotics foundation models often focus narrowly on control tasks, neglecting to assess their capabilities in understanding and generating content across individual modalities. Multinet's comprehensive evaluation framework highlights opportunities to enhance these models' fundamental capabilities, paving the way for more powerful, truly generalist robotic systems.

Standardizing Robotics Data A significant contribution of this work is our open-source toolkit for standardizing robotics and reinforcement learning data. Many existing datasets suffer from outdated formats, poor maintenance, and accessibility issues. Our toolkit addresses these challenges by:

- Providing stable access methods for diverse RL and robotics datasets
- Converting various data formats to a unified TensorFlow dataset format

- Enabling easy local storage and usage for training, fine-tuning, and evaluation

Fostering Advanced Evaluation Methods While we currently use Mean Squared Error between predicted and ground-truth actions to evaluate performance on the OpenX-Embodiment dataset, we acknowledge this metric's limitations. By highlighting the need for more sophisticated evaluation methods, we aim to encourage:

- Development of more advanced simulation environments
- Creation of evaluation methods that better approximate real-world performance
- Innovation in offline RL and robotics task assessment

Through these contributions, Multinet not only provides immediate value for evaluating and developing generalist AI systems but also highlights critical areas for future research and development in the field.

6 Future Work

The release of Multinet marks an important first step toward a new paradigm of foundation models, but significant opportunities remain for expansion and improvement. Our vision for future iterations of Multinet encompasses several ambitious directions.

A key priority is deepening our understanding of how control-task training affects model capabilities. While current VLAs show promising results on control tasks, we plan to systematically evaluate their performance on pure vision-language and language tasks to assess whether fine-tuning or co-training on control tasks compromises their capabilities in individual modalities. This investigation will provide crucial insights into the trade-offs involved in developing truly generalist models.

We also aim to broaden our evaluation scope beyond the OpenX-Embodiment dataset. By incorporating the diverse control tasks described in this paper, we can better understand how VLAs and generalist models perform on completely out-of-distribution data. This expansion will help identify both the strengths and limitations of current approaches while suggesting directions for improvement.

While our current profiling efforts focus on zero-shot performance, future work will explore few-shot learning and fine-tuning scenarios. Of particular interest is the potential transfer of VLA capabilities to novel domains. For instance, we are exploring how these models might be adapted to software environments, potentially enabling more capable digital agents by leveraging insights from embodied learning.

Finally, we envision transforming Multinet from its current offline form into an online benchmark. This evolution may include the development of simulation environments for both 2D and 3D control tasks, enabling more dynamic and interactive evaluation of model capabilities. Such an extension would provide richer insights into model behavior and better approximate real-world deployment scenarios.

Through these future developments, we aim to establish Multinet as a comprehensive and evolving platform for advancing the field of generalist AI systems. By providing increasingly sophisticated tools and benchmarks, we hope to accelerate progress toward more capable and versatile artificial intelligence.

7 References

References